EURASIP Journal on Applied Signal Processing 2003:11, 1064–1073

An Integrated Real-Time Beamforming and Postfiltering System for Nonstationary Noise Environments

Israel Cohen
Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel
Email: icohen@ee.technion.ac.il

Sharon Gannot
School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel
Email: gannot@siglab.technion.ac.il

Baruch Berdugo
Lamar Signal Processing, Ltd., Andrea Electronics Corp., P.O. Box 573, Yokneam Ilit 20692, Israel
Email: bberdugo@lamar.co.il

Received 1 September 2002 and in revised form 6 March 2003

We present a novel approach for real-time multichannel speech enhancement in environments of nonstationary noise and time-varying acoustical transfer functions (ATFs). The proposed system integrates adaptive beamforming, ATF identification, soft signal detection, and multichannel postfiltering. The noise canceller branch of the beamformer and the ATF identification are adaptively updated online, based on hypothesis test results. The noise canceller is updated only during stationary noise frames, and the ATF identification is carried out only when desired source components have been detected. The hypothesis testing is based on the nonstationarity of the signals and the transient power ratio between the beamformer primary output and its reference noise signals. Following the beamforming and the hypothesis testing, estimates for the signal presence probability and for the noise power spectral density are derived. Subsequently, an optimal spectral gain function that minimizes the mean square error of the log-spectral amplitude (LSA) is applied. Experimental results demonstrate the usefulness of the proposed system in nonstationary noise environments.

Keywords and phrases: array signal processing, signal detection, acoustic noise measurement, speech enhancement, spectral analysis, adaptive signal processing.

1. INTRODUCTION

Postfiltering methods for multimicrophone speech enhancement algorithms have recently attracted an increased interest. It is well known that beamforming methods yield a significant improvement in speech quality [1]. However, when the noise field is spatially incoherent or diffuse, the noise reduction is insufficient and additional postfiltering is normally required [2]. Most multimicrophone speech enhancement methods comprise a multichannel part (either delay-sum beamformer or generalized sidelobe canceller (GSC) [3]) followed by a postfilter, which is based on Wiener filtering (sometimes in conjunction with spectral subtraction). Numerous articles have been published on that subject, for example, [4, 5, 6, 7, 8, 9, 10, 11, 12] to mention just a few. A major drawback of these multichannel postfiltering techniques is that highly nonstationary noise components are not dealt with. The time variation of the interfering signals is assumed to be sufficiently slow such that the postfilter can track and adapt to the changes in the noise statistics. Unfortunately, transient interferences are often much too brief and abrupt for the conventional tracking methods.

Recently, a multichannel postfilter was incorporated into the GSC beamformer [13, 14]. The use of both the beamformer primary output and the reference noise signals (resulting from the blocking branch of the GSC) for distinguishing between desired speech transients and interfering transients enables the algorithm to work in nonstationary noise environments. In [15], the multichannel postfilter is combined with the transfer function GSC (TF GSC) [16], and compared with single-microphone postfilters, namely, the mixture-maximum (MIXMAX) [17] and the optimally modified log-spectral amplitude (OM LSA) estimator [18]. The multichannel postfilter, combined with the TF GSC, proved the best for handling abrupt noise spectral variations. However, in all past contributions the beamformer

stage feeds the postfilter but the adverse is not true. The decisions made by the postfilter, distinguishing between speech, stationary noise, and transient noise, might be fed back to the beamformer to enable the use of the method in real-time applications. Exploiting this information will also enable the tracking of the acoustical transfer functions (ATFs), caused by talker movements.

In this paper, we present a real-time multichannel speech enhancement system, which integrates adaptive beamforming and multichannel postfiltering. The beamformer is based on the TF GSC. However, the requirement for the stationarity of the noise is relaxed. Furthermore, we allow the ATFs to vary in time, which entails an online system identification procedure. We define hypotheses that indicate either the absence of transients, presence of an interfering transient, or presence of desired source components (the stationary noise persists in all cases). The noise canceller branch of the beamformer is updated only during the absence of transients, and the ATF identification is carried out only when desired source components are present. Following the beamforming and the hypothesis testing, estimates for the signal presence probability and for the noise power spectral density (PSD) are derived. Subsequently, an optimal spectral gain function that minimizes the mean square error of the log-spectral amplitude (LSA) is applied.

The performance of the proposed system is evaluated under nonstationary noise conditions, and compared to that obtained with a single-channel postfiltering approach. We show that single-channel postfiltering is inefficient at attenuating highly nonstationary noise components since it lacks the ability to differentiate such components from the desired source components. By contrast, the proposed system achieves a significantly reduced level of background noise, whether stationary or not, without further distorting the signal components.

The paper is organized as follows. In Section 2, we introduce a novel approach for real-time beamforming in nonstationary noise environments, under the circumstances of time-varying ATFs. The noise canceller branch of the beamformer and the ATF identification are adaptively updated online, based on hypothesis test results. In Section 3, the problem of hypothesis testing in the time-frequency plane is addressed. Signal components are detected and discriminated from the transient noise components based on the transient power ratio between the beamformer primary output and its reference noise signals. In Section 4, we introduce the multichannel postfilter and outline the implementation steps of the integrated TF GSC and multichannel postfiltering algorithm. Finally, in Section 5, we evaluate the proposed system and present experimental results which validate its usefulness.

2. TRANSFER FUNCTION GENERALIZED SIDELOBE CANCELLING

Let $x(t)$ denote a desired speech source signal that, subject to some acoustic propagation, is received by M microphones along with additive uncorrelated interfering signals. The interference at the ith sensor comprises a pseudostationary noise signal $d_{is}(t)$ and a transient noise component $d_{it}(t)$. The observed signals are given by

$$z_i(t) = a_i(t) * x(t) + d_{is}(t) + d_{it}(t), \quad i = 1, \ldots, M, \tag{1}$$

where $a_i(t)$ is the impulse response of the ith sensor to the desired source and $*$ denotes convolution. Using the short-time Fourier transform (STFT), we have

$$\mathbf{Z}(k,\ell) = \mathbf{A}(k,\ell)X(k,\ell) + \mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell) \tag{2}$$

in the time-frequency domain, where k represents the frequency bin index, ℓ the frame index, and

$$\begin{aligned}
\mathbf{Z}(k,\ell) &\triangleq \begin{bmatrix} Z_1(k,\ell) & Z_2(k,\ell) & \cdots & Z_M(k,\ell) \end{bmatrix}^T, \\
\mathbf{A}(k,\ell) &\triangleq \begin{bmatrix} A_1(k,\ell) & A_2(k,\ell) & \cdots & A_M(k,\ell) \end{bmatrix}^T, \\
\mathbf{D}_s(k,\ell) &\triangleq \begin{bmatrix} D_{1s}(k,\ell) & D_{2s}(k,\ell) & \cdots & D_{Ms}(k,\ell) \end{bmatrix}^T, \\
\mathbf{D}_t(k,\ell) &\triangleq \begin{bmatrix} D_{1t}(k,\ell) & D_{2t}(k,\ell) & \cdots & D_{Mt}(k,\ell) \end{bmatrix}^T.
\end{aligned} \tag{3}$$

The observed noisy signals are processed by the system shown in Figure 1. This structure is a modification to the recently proposed TF GSC [16], which is an extension of the linearly constrained adaptive beamformer [3, 19] for arbitrary ATFs, $\mathbf{A}(k,\ell)$. In [16], transient interferences are not dealt with since signal enhancement is based on the nonstationarity of the desired source signal, contrasted with the stationarity of the noise signal. As such, the ATF estimation was conducted in an offline manner. Here, the requirement for the stationarity of the noise is relaxed. So a mechanism for discriminating interfering transients from desired signal components must be included. Furthermore, in contrast to the assumption of time-invariant ATFs in [16], we allow time-varying ATFs provided that their change rate is slow in comparison to that of the speech statistics. This entails online adaptive estimates for the ATFs.

The beamformer comprises three parts: a fixed beamformer $\mathbf{W}$, which aligns the desired signal components; a blocking matrix $\mathbf{B}$, which blocks the desired components, thus yielding the reference noise signals $\{U_i : 2 \le i \le M\}$; and a multichannel adaptive noise canceller $\{H_i : 2 \le i \le M\}$, which eliminates the stationary noise that leaks through the sidelobes of the fixed beamformer. The reference noise signals $\mathbf{U}(k,\ell) = [U_2(k,\ell) \;\; U_3(k,\ell) \;\; \cdots \;\; U_M(k,\ell)]^T$ are generated by applying the blocking matrix to the observed signal vector:

$$\begin{aligned}
\mathbf{U}(k,\ell) &= \mathbf{B}^H(k,\ell)\mathbf{Z}(k,\ell) \\
&= \mathbf{B}^H(k,\ell)[\mathbf{A}(k,\ell)X(k,\ell) + \mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell)].
\end{aligned} \tag{4}$$

The reference noise signals are emphasized by the adaptive noise canceller and subtracted from the output of the fixed beamformer, yielding

$$Y(k,\ell) = [\mathbf{W}^H(k,\ell) - \mathbf{H}^H(k,\ell)\mathbf{B}^H(k,\ell)]\mathbf{Z}(k,\ell), \tag{5}$$

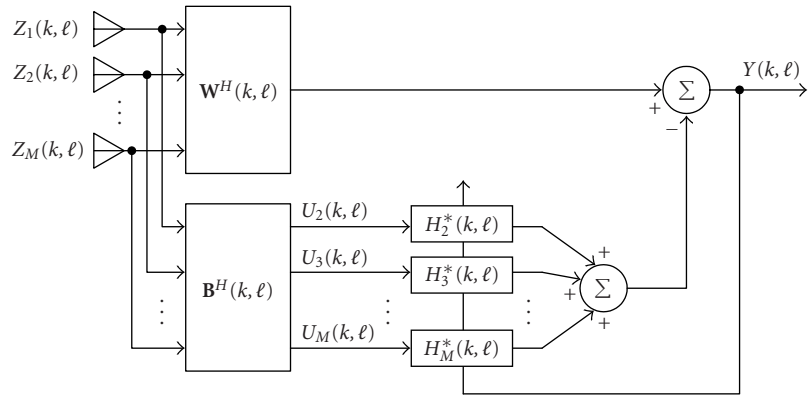

Figure 1: Block diagram of the TF GSC.

where $\mathbf{H}(k,\ell) = [H_2(k,\ell)\ H_3(k,\ell)\ \cdots\ H_M(k,\ell)]^T$. It is worth mentioning that a perfect blocking matrix implies $\mathbf{B}^H(k,\ell)\mathbf{A}(k,\ell) = 0$. In that case, $\mathbf{U}(k,\ell)$ indeed contains only noise components:

$$\mathbf{U}(k,\ell) = \mathbf{B}^H(k,\ell)\big[\mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell)\big]. \tag{6}$$

In general, however, $\mathbf{B}^H(k,\ell)\mathbf{A}(k,\ell) \neq 0$, thus desired signal components may leak into the noise reference signals.

Let three hypotheses H_{0s}, H_{0t}, and H_1 indicate, respectively, the absence of transients, presence of an interfering transient, and presence of a desired source transient at the beamformer output. The optimal solution for the filters $\mathbf{H}(k,\ell)$ is obtained by minimizing the power of the beamformer output during the stationary noise frames (i.e., when H_{0s} is true) [20]. Let $\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell) = E\{\mathbf{D}_s(k,\ell)\mathbf{D}_s^H(k,\ell)\}$ denote the PSD matrix of the input stationary noise. Then, the power of the stationary noise at the beamformer output is minimized by solving the unconstrained optimization problem

$$\begin{aligned}\min_{\mathbf{H}} \Big\{&\big[\mathbf{W}(k,\ell) - \mathbf{B}(k,\ell)\mathbf{H}(k,\ell)\big]^H \mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell)\\ &\times \big[\mathbf{W}(k,\ell) - \mathbf{B}(k,\ell)\mathbf{H}(k,\ell)\big]\Big\}.\end{aligned} \tag{7}$$

A multichannel Wiener solution is given by [21]

$$\begin{aligned}\mathbf{H}(k,\ell) = &\big[\mathbf{B}^H(k,\ell)\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell)\mathbf{B}(k)\big]^{-1}\\ &\times \mathbf{B}^H(k,\ell)\mathbf{\Phi}_{\mathbf{D}_s\mathbf{D}_s}(k,\ell)\mathbf{W}(k,\ell).\end{aligned} \tag{8}$$

In practice, this optimization problem is solved by using the normalized least mean squares (LMS) algorithm [20]

$$\mathbf{H}(k,\ell+1) = \begin{cases}\mathbf{H}(k,\ell) + \dfrac{\mu_h}{P_{\mathrm{est}}(k,\ell)}\mathbf{U}(k,\ell)Y^*(k,\ell), & \text{if } \mathrm{H}_{0s} \text{ is true},\\ \mathbf{H}(k,\ell), & \text{otherwise},\end{cases} \tag{9}$$

where

$$P_{\mathrm{est}}(k,\ell) = \begin{cases}\alpha_p P_{\mathrm{est}}(k,\ell-1) + (1-\alpha_p)\big\|\mathbf{U}(k,\ell)\big\|^2, & \text{if } \mathrm{H}_{0s} \text{ is true},\\ P_{\mathrm{est}}(k,\ell-1), & \text{otherwise},\end{cases} \tag{10}$$

represents the power of the noise reference signals, μ_h is a step factor that regulates the convergence rate, and α_p is a smoothing parameter.

The fixed beamformer implements the alignment of the desired signal by applying a matched filter to the ATF ratios [16]:

$$\mathbf{W}(k,\ell) \triangleq \frac{\tilde{\mathbf{A}}(k,\ell)}{\big\|\tilde{\mathbf{A}}(k,\ell)\big\|^2}, \tag{11}$$

where

$$\begin{aligned}\tilde{\mathbf{A}}(k,\ell) &\triangleq \frac{\mathbf{A}(k,\ell)}{A_1(k,\ell)}\\ &= \left[1\ \ \frac{A_2(k,\ell)}{A_1(k,\ell)}\ \ \cdots\ \ \frac{A_M(k,\ell)}{A_1(k)}\right]^T\\ &\triangleq \left[1\ \ \tilde{A}_2(k,\ell)\ \ \cdots\ \ \tilde{A}_M(k,\ell)\right]^T\end{aligned} \tag{12}$$

denotes ATF ratios, with $A_1(k,\ell)$ chosen arbitrarily as the reference ATF. The blocking matrix $\mathbf{B}$ is aimed at eliminating the desired signal and constructing reference noise signals. A proper (but not unique) choice of the blocking matrix is given by [16]

$$\mathbf{B}(k,\ell) = \begin{bmatrix}-\tilde{A}_2^*(k,\ell) & -\tilde{A}_3^*(k,\ell) & \cdots & -\tilde{A}_M^*(k,\ell)\\ 1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}. \tag{13}$$

Hence, for implementing both the fixed beamformer and the

blocking matrix, we need to estimate the ATF ratios. In contrast to previous works [14, 15, 16], the system identification should be incorporated into the adaptive procedure since the ATFs are time varying. In [16], the system identification procedure is based on the nonstationarity of the desired signal. Here, a modified version is introduced, employing the already available time-frequency analysis of the beamformer and the decisions made by hypothesis testing.

From (4) and (13), we have the following input-output relation between $Z_1(k,\ell)$ and $Z_i(k,\ell)$:

$$Z_i(k,\ell) = \tilde{A}_i(k,\ell)Z_1(k,\ell) + U_i(k,\ell), \quad i = 2,\ldots,M. \tag{14}$$

Accordingly,

$$\begin{aligned}&\phi_{Z_iZ_1}(k,\ell)\\ &\quad = \tilde{A}_i(k,\ell)\phi_{Z_1Z_1}(k,\ell) + \phi_{U_iZ_1}(k,\ell), \quad i = 2,\ldots,M,\end{aligned} \tag{15}$$

where $\phi_{Z_iZ_1}(k,\ell) = E\{Z_i(k,\ell)Z_1^*(k,\ell)\}$ is the cross PSD between $z_i(t)$ and $z_1(t)$, and $\phi_{U_iZ_1}(k,\ell)$ is the cross PSD between $u_i(t)$ and $z_1(t)$. The use of standard system identification methods is inapplicable since the interference signal $u_i(t)$ is strongly correlated to the system input $z_1(t)$. However, when hypothesis H_1 is true, that is, when transient noise is absent, the cross PSD $\phi_{U_iZ_1}(k,\ell)$ becomes stationary. Therefore, $\phi_{U_iZ_1}(k,\ell)$ may be replaced with $\phi_{U_iZ_1}(k)$.

For estimating the ATF ratios $\tilde{\mathbf{A}}(k,\ell)$, we need to collect several estimates of the PSD $\boldsymbol{\phi}_{\mathbf{Z}Z_1}(k,\ell)$, each of which is based on averaging several frames. Let a segment define a concatenation of N frames for which the hypothesis H_1 is true, and let an interval contain R such segments. Then, the PSD estimation in each segment r ($r = 1,\ldots,R$) is obtained by averaging the periodograms over N frames:

$$\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(r)}(k,\ell) = \frac{1}{N}\sum_{\ell\in\mathcal{L}_r}\mathbf{Z}(k,\ell)Z_1^*(k,\ell), \tag{16}$$

where $\mathcal{L}_r$ represents the set of frames that belong to the rth segment. Denoting by $\varepsilon_i^{(r)}(k,\ell) = \hat{\phi}_{U_iZ_1}^{(r)}(k,\ell) - \phi_{U_iZ_1}(k)$ the estimation error of the cross PSD between $u_i(t)$ and $z_1(t)$ in the rth segment, (15) implies that

$$\begin{aligned}\hat{\phi}_{Z_iZ_1}^{(r)}(k,\ell) = \tilde{A}_i(k,\ell)\hat{\phi}_{Z_1Z_1}^{(r)}(k,\ell) + \phi_{U_iZ_1}(k) + \varepsilon_i^{(r)}(k,\ell),\\ i = 2,\ldots,M,\ r = 1,2,\ldots,R.\end{aligned} \tag{17}$$

The least squares (LS) solution to this overdetermined set of equation is given by [16]

$$\tilde{\mathbf{A}}(k,\ell) = \frac{\langle\hat{\phi}_{Z_1Z_1}(k,\ell)\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}(k,\ell)\rangle - \langle\hat{\phi}_{Z_1Z_1}(k,\ell)\rangle\langle\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}(k,\ell)\rangle}{\langle\hat{\phi}_{Z_1Z_1}^2(k,\ell)\rangle - \langle\hat{\phi}_{Z_1Z_1}(k,\ell)\rangle^2}, \tag{18}$$

where the average operation on $\beta(k,\ell)$ is defined by

$$\langle\beta(k,\ell)\rangle \triangleq \frac{1}{R}\sum_{r=1}^{R}\beta^{(r)}(k,\ell). \tag{19}$$

Practically, the estimates for $\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(r)}(k,\ell)$ ($r = 1,\ldots,R$) are recursively obtained as follows. In each time-frequency bin (k,ℓ), we assume that R PSD estimates are already available (excluding initial conditions). Values of $\tilde{\mathbf{A}}(k,\ell)$ are thus ready for use in the next frame $(k,\ell+1)$. Frames for which hypothesis H_1 is true are collected for obtaining a new PSD estimate $\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(R+1)}(k,\ell)$:

$$\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(R+1)}(k,\ell+1) = \hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(R+1)}(k,\ell) + \frac{1}{N}\mathbf{Z}(k,\ell)Z_1^*(k,\ell). \tag{20}$$

A counter n_k is employed for counting the number of times (20) is processed (counting the number of H_1 frames in frequency bin k). Whenever n_k reaches N, the estimate in segment $R+1$ is stacked into the previous estimates, the oldest estimate ($r = 1$) is discarded, and n_k is initialized. The new R estimates are then used for obtaining a new estimate for the ATF ratios $\tilde{\mathbf{A}}(k,\ell+1)$ for the next bin $(k,\ell+1)$. This procedure is active for all frames ℓ enabling a real-time tracking of the beamformer.

Altogether, an interval containing $N \times R$ frames, for which H_1 is true, is used for obtaining an estimate for $\tilde{\mathbf{A}}(k,\ell)$. Special attention should be given for choosing this quantity. On the one hand, it should be long enough for stabilizing the solution. On the other hand, it should be short enough for the ATF quasistationarity assumption to hold during the interval. We note that for frequency bins with low speech content, the interval (observation time) required for obtaining an estimate for $\tilde{\mathbf{A}}(k,\ell)$ might be very long, since only frames for which H_1 is true are collected.

3. HYPOTHESIS TESTING

Generally, the TF GSC output comprises three components: a nonstationary desired source component, a pseudostationary noise component, and a transient interference. Our objective is to determine which category a given time-frequency bin belongs to, based on the beamformer output and the reference signals. Clearly, if transients have not been detected at the beamformer output and the reference signals, we can accept hypothesis H_{0s}. In case a transient is detected at the beamformer output, but not at the reference signals, the transient is likely a source component, and therefore we determine that H_1 is true. On the contrary, a transient that is detected at one of the reference signals but not at the beamformer output is likely an interfering component, which implies that H_{0t} is true. In case a transient is simultaneously detected at the beamformer output and at one of the reference signals, a further test is required, which involves the ratio between the transient power at beamformer output and the transient power at the reference signals.

Let $\mathcal{S}$ be a smoothing operator in the PSD

$$\begin{aligned}\mathcal{S}Y(k,\ell) = \alpha_s\cdot\mathcal{S}Y(k,\ell-1)\\ + (1-\alpha_s)\sum_{i=-w}^{w}b_i\left|Y(k-i,\ell)\right|^2,\end{aligned} \tag{21}$$

where α_s ($0 \le \alpha_s \le 1$) is a forgetting factor for the smoothing

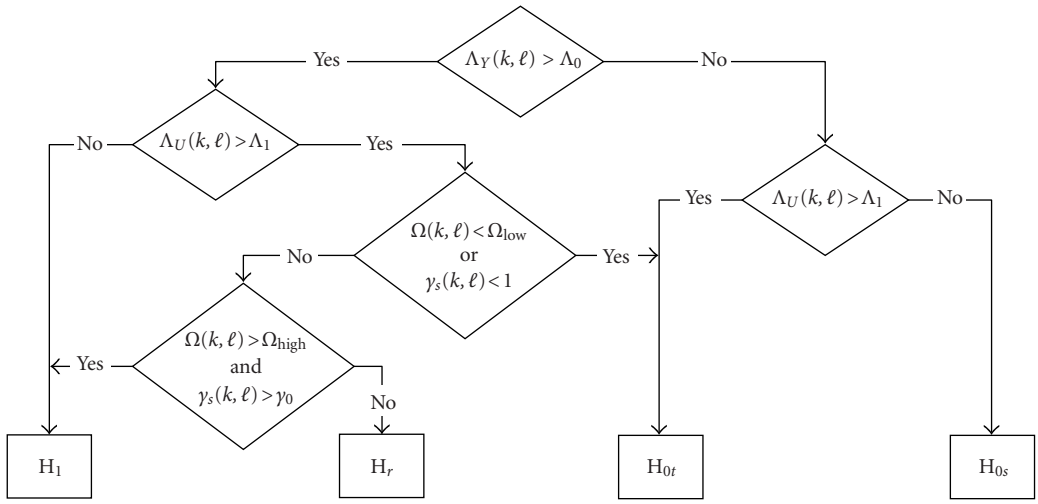

Figure 2: Block diagram for the hypothesis testing.

in time, and b is a normalized window function ($\sum_{i=-w}^{w} b_i = 1$) that determines the order of smoothing in frequency. Let $\mathcal{M}$ denote an estimator for the PSD of the background pseudostationary noise, derived using the *minima controlled recursive averaging* approach [18, 22]. The decision rules for detecting transients at the TF GSC output and reference signals are

$$\Lambda_Y(k,\ell) \triangleq \frac{\mathcal{S}Y(k,\ell)}{\mathcal{M}Y(k,\ell)} > \Lambda_0, \tag{22}$$

$$\Lambda_{\mathbf{U}}(k,\ell) \triangleq \max_{2\le i\le M}\left\{\frac{\mathcal{S}U_i(k,\ell)}{\mathcal{M}U_i(k,\ell)}\right\} > \Lambda_1, \tag{23}$$

respectively, where Λ_Y and Λ_U denote measures of the local nonstationarities (LNS), and Λ_0 and Λ_1 are the corresponding threshold values for detecting transients [14]. The transient beam-to-reference ratio (TBRR) is defined by the ratio between the transient power of the beamformer output and the transient power of the strongest reference signal:

$$\Omega(k,\ell) = \frac{\mathcal{S}Y(k,\ell) - \mathcal{M}Y(k,\ell)}{\max_{2\le i\le M}\{\mathcal{S}U_i(k,\ell) - \mathcal{M}U_i(k,\ell)\}}. \tag{24}$$

Transient signal components are relatively strong at the beamformer output, whereas transient noise components are relatively strong at one of the reference signals. Hence, we expect $\Omega(k,\ell)$ to be large for signal transients and small for noise transients. Assuming that there exist thresholds $\Omega_{\text{high}}(k)$ and $\Omega_{\text{low}}(k)$ such that

$$\Omega(k,\ell)|_{\mathrm{H}_{0t}} \le \Omega_{\text{low}}(k) \le \Omega_{\text{high}}(k) \le \Omega(k,\ell)|_{\mathrm{H}_1}, \tag{25}$$

the decision rule for differentiating desired signal components from the transient interference components is

$$\begin{aligned} &\mathrm{H}_{0t} : \gamma_s(k,\ell) \le 1 \text{ or } \Omega(k,\ell) \le \Omega_{\text{low}}(k),\\ &\mathrm{H}_1 : \gamma_s(k,\ell) \ge \gamma_0 \text{ and } \Omega(k,\ell) \ge \Omega_{\text{high}}(k),\\ &\mathrm{H}_r : \text{otherwise,} \end{aligned} \tag{26}$$

where

$$\gamma_s(k,\ell) \triangleq \frac{|Y(k,\ell)|^2}{\mathcal{M}Y(k,\ell)} \tag{27}$$

represents the a posteriori SNR at the beamformer output with respect to the pseudostationary noise, γ_0 denotes a constant satisfying $\mathcal{P}(\gamma_s(k,\ell) \ge \gamma_0|\mathrm{H}_{0s}) < \epsilon$ for a certain significance level ϵ, and H_r designates a *reject* option where the conditional error of making a decision between H_{0t} and H_1 is high.

Figure 2 summarizes a block diagram for the hypothesis testing. The hypothesis testing is carried out in the time-frequency plane for each frame and frequency bin. Hypothesis H_{0s} is accepted when transients have been detected neither at the beamformer output nor at the reference signals. In case a transient is detected at the beamformer output but not at the reference signals, we accept H_1. On the other hand, if a transient is detected at one of the reference signals but not at the beamformer output, we accept H_{0t}. In case a transient is detected simultaneously at the beamformer output and at one of the reference signals, we compute the TBRR $\Omega(k,\ell)$ and the a posteriori SNR at the beamformer output with respect to the pseudostationary noise $\gamma_s(k,\ell)$, and decide on the hypothesis according to (26).

4. MULTICHANNEL POSTFILTERING

In this section, we address the problem of estimating the time-varying PSD of the TF GSC output noise and present the multichannel postfiltering technique. Figure 3 describes a block diagram of the multichannel postfiltering. Following the hypothesis testing, an estimate $\hat{q}(k,\ell)$ for the a priori signal absence probability is produced. Subsequently, we derive an estimate $p(k,\ell) \triangleq \mathcal{P}(\mathrm{H}_1|Y,\mathbf{U})$ for the signal presence probability and an estimate $\hat{\lambda}_d(k,\ell)$ for the noise PSD.

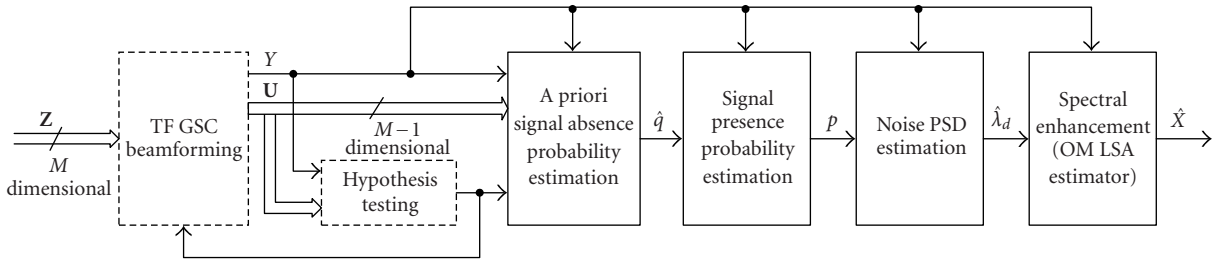

Figure 3: Block diagram of the multichannel postfiltering.

Finally, spectral enhancement of the beamformer output is achieved by applying the OM LSA gain function [18], which minimizes the mean square error of the LSA under signal presence uncertainty.

Based on a Gaussian statistical model [23], the signal presence probability is given by

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}\left(1+\xi(k,\ell)\right)\exp\left(-\upsilon(k,\ell)\right)\right\}^{-1}, \tag{28}$$

where $\xi(k,\ell) \triangleq \lambda_x(k,\ell)/\lambda_d(k,\ell)$ is the a priori SNR, $\lambda_d(k,\ell)$ is the noise PSD at the beamformer output, $\upsilon(k,\ell) \triangleq \gamma(k,\ell)\xi(k,\ell)/(1+\xi(k,\ell))$, and $\gamma(k,\ell) \triangleq |Y(k,\ell)|^2/\lambda_d(k,\ell)$ is the a posteriori SNR. The a priori signal absence probability $\hat{q}(k,\ell)$ is set to 1 if signal absence hypotheses (H_{0s} or H_{0t}) are accepted and is set to 0 if signal presence hypothesis (H_1) is accepted. In case of the reject hypothesis H_r, a soft signal detection is accomplished by letting $\hat{q}(k,\ell)$ be inversely proportional to $\Omega(k,\ell)$ and $\gamma_s(k,\ell)$:

$$\hat{q}(k,\ell) = \max\left\{\frac{\gamma_0 - \gamma_s(k,\ell)}{\gamma_0 - 1}, \frac{\Omega_{\text{high}} - \Omega(k,\ell)}{\Omega_{\text{high}} - \Omega_{\text{low}}}\right\}. \tag{29}$$

The a priori SNR is estimated by [18]

$$\begin{aligned}\hat{\xi}(k,\ell) = {} & \alpha G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1) \\ & + (1-\alpha)\max\{\gamma(k,\ell)-1, 0\},\end{aligned} \tag{30}$$

where α is a weighting factor that controls the trade-off between noise reduction and signal distortion, and

$$G_{H_1}(k,\ell) \triangleq \frac{\xi(k,\ell)}{1+\xi(k,\ell)}\exp\left(\frac{1}{2}\int_{\upsilon(k,\ell)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{31}$$

is the spectral gain function of the LSA estimator when the signal is surely present [24]. An estimate for noise PSD is obtained by recursively averaging past spectral power values of the noisy measurement, using a time-varying frequency-dependent smoothing parameter. The recursive averaging is given by

$$\begin{aligned}\hat{\lambda}_d(k,\ell+1) = {} & \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) \\ & + \beta[1-\tilde{\alpha}_d(k,\ell)]\,|Y(k,\ell)|^2,\end{aligned} \tag{32}$$

where the smoothing parameter $\tilde{\alpha}_d(k,\ell)$ is determined by the signal presence probability $p(k,\ell)$:

$$\tilde{\alpha}_d(k,\ell) \triangleq \alpha_d + (1-\alpha_d)\,p(k,\ell), \tag{33}$$

and β is a factor that compensates the bias when the signal is absent. The constant α_d ($0 < \alpha_d < 1$) represents the minimal smoothing parameter value. The smoothing parameter is close to 1 when the signal is present to prevent an increase in the noise estimate as a result of signal components. It decreases when the probability of signal presence decreases to allow a fast update of the noise estimate.

The estimate of the clean signal STFT is finally given by

$$\hat{X}(k,\ell) = G(k,\ell)Y(k,\ell), \tag{34}$$

where

$$G(k,\ell) = \{G_{H_1}(k,\ell)\}^{p(k,\ell)}G_{\min}^{1-p(k,\ell)} \tag{35}$$

is the OM LSA gain function and $G_{\min}$ denotes a lower bound constraint for the gain when the signal is absent. The implementation of the integrated TF GSC and multichannel postfiltering algorithm is summarized in Algorithm 1. Typical values of the respective parameters, for a sampling rate of 8 kHz, are given in Table 1. The STFT and its inverse are implemented with biorthogonal Hamming windows of 256 samples length (32 milliseconds) and 64 samples frame update step (75% overlap between successive windows).

5. EXPERIMENTAL RESULTS

In this section, we compare under nonstationary noise conditions the performance of the proposed real-time system to an offline system consisting of a TF GSC and a single-channel postfilter. The performance evaluation includes objective quality measures, a subjective study of speech spectrograms, and informal listening tests.

A linear array, consisting of four microphones with 5 cm spacing is mounted in a car on the visor. Clean speech signals are recorded at a sampling rate of 8 kHz in the absence of background noise (standing car, silent environment). An interfering speaker and car noise signals are recorded while the car speed is about 60 km/h, and the window next to the driver is slightly open (about 5 cm; the other windows are

Initialize variables at the first frame for all frequency bins k:

$G_{\mathrm{H}_1}(k,0) = \gamma(k,0) = 1$; $P_{\mathrm{est}}(k,0) = \|\mathbf{U}(k,0)\|^2$;

$\mathcal{S}Y(k,0) = \mathcal{M}Y(k,0) = \hat{\lambda}_d(k,0) = |Y(k,0)|^2$;

Let $n_k = 0$; % n_k is a counter for H_1 frames in frequency bin k.

For $i = 2,\ldots,M$,

$\mathcal{S}U_i(k,0) = \mathcal{M}U_i(k,0) = |U_i(k,0)|^2$; $H_i(k,0) = 0$; $\tilde{A}_i(k,0) = 1$.

For all time frames ℓ

For all frequency bins k

Compute the reference noise signals $\mathbf{U}(k,\ell)$ using (4), and the TF GSC output $Y(k,\ell)$ using (5).

Compute the recursively averaged spectrum of the TF GSC output and reference signals, $\mathcal{S}Y(k,\ell)$ and $\mathcal{S}U_i(k,\ell)$, using (21), and update the MCRA estimates of the background pseudostationary noise $\mathcal{M}Y(k,\ell)$ and $\mathcal{M}U_i(k,\ell)$ $(i = 2,\ldots,M)$ using [22].

Compute the local nonstationarities of the TF GSC output and reference signals $\Lambda_Y(k,\ell)$ and $\Lambda_{\mathbf{U}}(k,\ell)$ using (22) and (23).

Using the block diagram for the hypothesis testing (Figure 2), determine the relevant hypothesis; it possibly requires computation of the transient beam-to-reference ratio $\Omega(k,\ell)$ using (24), and the a posteriori SNR at the beamformer output with respect to the pseudostationary noise $\gamma_s(k,\ell)$ using (27).

Update the estimate for the power of the reference signals $P_{\mathrm{est}}(k,\ell)$ using (10). In case of absence of transients (H_{0s}), update the multichannel adaptive noise canceller $\mathbf{H}(k,\ell+1)$ using (9).

In case of desired signal presence (H_1), update the estimate $\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(R+1)}(k,\ell+1)$ using (20), and increment n_k by 1.

If $n_k \equiv N$, then store $\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(r+1)}(k,\ell+1)$ as $\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(r)}(k,\ell+1)$ for $r = 1,\ldots,R$, update the ATF ratios $\tilde{\mathbf{A}}(k,\ell)$ using (18), and reset $\hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}^{(R+1)}(k,\ell+1)$ and n_k to zero.

In case of H_{0s} or H_{0t}, set the a priori signal absence probability $\hat{q}(k,\ell)$ to 1. In case of H_1, set $\hat{q}(k,\ell)$ to 0. In case of H_r, compute $\hat{q}(k,\ell)$ according to (29).

Compute the a priori SNR $\hat{\xi}(k,\ell)$ using (30), the conditional gain $G_{\mathrm{H}_1}(k,\ell)$ using (31), and the signal presence probability $p(k,\ell)$ using (28).

Compute the time-varying smoothing parameter $\tilde{\alpha}_d(k,\ell)$ using (33) and update the noise spectrum estimate $\hat{\lambda}_d(k,\ell+1)$ using (32).

Compute the OM LSA estimate of the clean signal $\hat{X}(k,\ell)$ using (34) and (35).

Algorithm 1: The integrated TF GSC and multichannel postfiltering algorithm.

Table 1: Values of parameters used in the implementation of the proposed algorithm for a sampling rate of 8 kHz.

Normalized LMS	$\alpha_p = 0.9$	$\mu_h = 0.05$
ATF identification	$N = 10$	$R = 10$
Hypothesis testing	$\alpha_s = 0.9$	$\gamma_0 = 4.6$
	$\Lambda_0 = 1.67$	$\Lambda_1 = 1.81$
	$\Omega_{\mathrm{low}} = 1$	$\Omega_{\mathrm{high}} = 3$
	$b = [0.25 \quad 0.5 \quad 0.25]$	
Noise PSD estimation	$\alpha_d = 0.85$	$\beta = 1.47$
Spectral enhancement	$\alpha = 0.92$	$G_{\mathrm{min}} = -20\,\mathrm{dB}$

closed). The input microphone signals are generated by mixing the speech and noise signals at various SNR levels in the range $[-5, 10]$ dB.

Offline TF GSC beamforming [16] is applied to the noisy multichannel signals, and its output is enhanced using the OM LSA estimator [18]. The result is referred to as single-channel postfiltering output. Alternatively, the proposed real-time integrated TF GSC and multichannel postfiltering is applied to the noisy signals. Its output is referred to as multichannel postfiltering output. Two objective quality measures are used. The first is segmental SNR, in dB, defined by [25]

$$\mathrm{SegSNR} = \frac{10}{L}\sum_{\ell=0}^{L-1} 10\log\frac{\sum_{n=0}^{K-1} x^2(n+\ell K/2)}{\sum_{n=0}^{K-1}\left[x(n+\ell K/2)-\hat{x}(n+\ell K/2)\right]^2}, \tag{36}$$

where L represents the number of frames in the signal, and $K = 256$ is the number of samples per frame (corresponding to 32 milliseconds frames, and 50% overlap). The SNR at each frame is limited to perceptually meaningful range between 35 dB and -10 dB [26, 27]. The second quality measure is log-spectral distance (LSD), in dB, which is defined by

$$\mathrm{LSD} = \frac{10}{L}\sum_{\ell=0}^{L-1}\left\{\frac{1}{K/2+1}\sum_{k=0}^{K/2}\left[\log\mathcal{C}X(k,\ell)-\log\mathcal{C}\hat{X}(k,\ell)\right]^2\right\}^{1/2}, \tag{37}$$

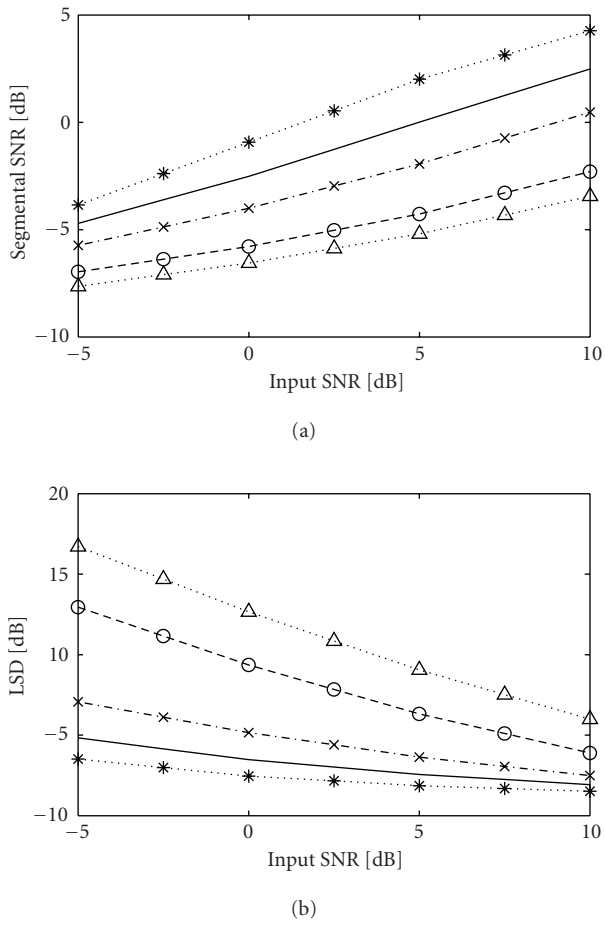

FIGURE 4: (a) Average segmental SNR and (b) average LSD at (△) microphone 1, (∘) TF GSC output, (×) single-channel postfiltering output, (solid line) multichannel postfiltering output, and (∗) theoretical limit postfiltering output.

where $\mathcal{C}X(k,\ell) \triangleq \max\{|X(k,\ell)|^2, \delta\}$ is the spectral power, clipped such that the log-spectral dynamic range is confined to about 50 dB (i.e., $\delta = 10^{-50/10}\max_{k,\ell}\{|X(k,\ell)|^2\}$).

Figure 4 shows experimental results obtained for various noise levels. The quality measures are evaluated at the first microphone, the offline TF GSC output, and the postfiltering outputs. A theoretical limit postfiltering, achievable by calculating the noise PSD from the noise itself, is also considered. It can be readily seen that TF GSC alone does not provide sufficient noise reduction in a car environment owing to its limited ability to reduce diffuse noise [16]. Furthermore, multichannel postfiltering is considerably better than single-channel postfiltering.

A subjective comparison between multichannel and single-channel postfiltering was conducted using speech spectrograms and validated by informal listening tests. Typical examples of speech spectrograms are presented in Figure 5. The noise PSD at the beamformer output varies substantially due to the residual interfering components of speech, wind blows, and passing cars. The TF GSC output is characterized by a high level of noise. Single-channel postfiltering suppresses pseudostationary noise components, but is inefficient at attenuating the transient noise components. By contrast, the proposed system achieves superior noise attenuation, while preserving the desired source components. This is verified by subjective informal listening tests.

6. CONCLUSION

We have described an integrated real-time beamforming and postfiltering system that is particularly advantageous in nonstationary noise environments. The system is based on the TF GSC beamformer and an OM LSA-based multichannel postfilter. The TF GSC beamformer primary output and the reference noise signals are exploited for deciding between speech, stationary noise, and transient noise hypotheses. The decisions are used for deriving estimators for the signal presence probability and for the noise PSD. The signal presence probability modifies the spectral gain function for estimating the clean signal spectral amplitude. It is worth mentioning that the postfilter is designed for suppressing the stationary noise as well as transient noise components that do not overlap with desired signal components in the time-frequency domain. The overlapping part between desired and undesired transients is not eliminated by the postfilter, to avoid signal distortion, particularly since such noise components are perceptually masked by the desired speech [28].

The proposed system was tested under nonstationary car noise conditions, and its performance was compared to that of a system based on single-channel postfiltering. While transient noise components are indistinguishable from desired source components when using a single-channel postfiltering approach, the enhancement of the beamformer output by multichannel postfiltering produces a significantly reduced level of residual transient noise without further distorting the desired signal components. We note that the computational complexity and practical simplifications of the proposed system were not addressed. Here, the main contribution is the incorporation of the hypothesis test results into the beamformer stage. The hypotheses control the noise canceller branch of the beamformer as well as the ATF identification, thus enabling real-time tracking of moving talkers.

The novel method has applications in realistic environments, where a desired speech signal is received by several microphones. In a typical office environment scenario, the speech signal is subject to propagation through time-varying ATFs (due to talker movements), stationary noise (e.g., air conditioner), and nonstationary interferences (e.g., radio or another talker). By adaptively updating the ATF ratios estimates, the TF GSC beamformer is consistently directed toward the desired speaker. An interfering source that is spatially separated from the desired source is therefore associated with TBRR lower than the desired source. Accordingly, transient noise components at the beamformer output can be differentiated from the desired speech components, and further suppressed by the postfilter.

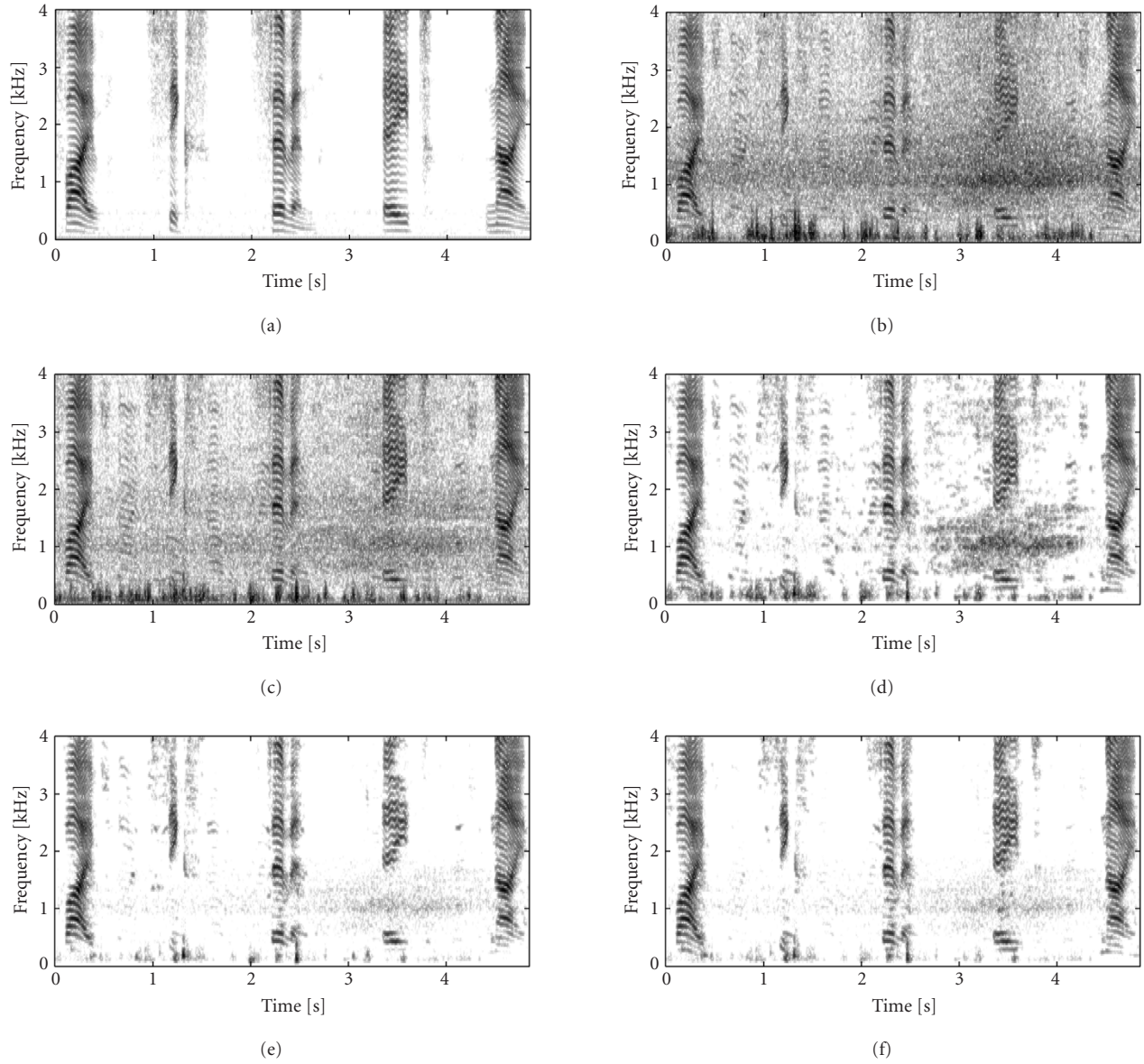

Figure 5: Speech spectrograms. (a) Original clean speech signal at microphone 1 (transcribed text: "five six seven eight nine"). (b) Noisy signal at microphone 1 (SNR = −0.9 dB, SegSNR = −6.2 dB, and LSD = 15.4 dB). (c) TF GSC output (SegSNR = −5.3 dB, LSD = 12.2 dB). (d) Single-channel postfiltering output (SegSNR = −3.8 dB, LSD = 7.4 dB). (e) Multichannel postfiltering output (SegSNR = −1.3 dB, LSD = 4.6 dB). (f) Theoretical limit (SegSNR = −0.4 dB, LSD = 4.0 dB).

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, Germany, 2001.

[2] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 3, pp. 39–60, Springer-Verlag, Berlin, Germany, 2001.

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[4] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. 13th IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2578–2581, New York, NY, USA, April 1988.

[5] R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post-filtering," *Electronics Letters*, vol. 26, no. 24, pp. 2036–2037, 1990.

[6] S. Fischer and K. U. Simmer, "An adaptive microphone array for hands-free communication," in *Proc. 4th International Workshop on Acoustic Echo and Noise Control*, pp. 44–47, Røros, Norway, June 1995.

[7] S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, no. 3-4, pp. 215–227, 1996.
[8] S. Fischer and K.-D. Kammeyer, "Broadband beamforming with adaptive post-filtering for speech acquisition in noisy environments," in *Proc. 22nd IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 359–362, Munich, Germany, April 1997.
[9] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. 22nd IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1167–1170, Munich, Germany, April 1997.
[10] K. U. Simmer, S. Fischer, and A. Wasiljeff, "Suppression of coherent and incoherent noise using a microphone array," *Annales des Télécommunications*, vol. 49, no. 7-8, pp. 439–446, 1994.
[11] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer," in *Proc. 6th International Workshop on Acoustic Echo and Noise Control*, pp. 100–103, Pocono Manor, Pa, USA, September 1999.
[12] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, vol. 34, no. 1-2, pp. 3–12, 2001.
[13] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. 27th IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 901–904, Orlando, Fla, USA, May 2002.
[14] I. Cohen, "Multi-channel post-filtering in non-stationary noise environments," to appear in IEEE Trans. Signal Processing.
[15] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and post-filtering," submitted to IEEE Trans. Speech and Audio Processing.
[16] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non-stationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.
[17] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.
[18] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
[19] C. W. Jim, "A comparison of two LMS constrained optimal array structures," *Proceedings of the IEEE*, vol. 65, no. 12, pp. 1730–1731, 1977.
[20] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1985.
[21] S. Nordholm, I. Claesson, and P. Eriksson, "The broadband Wiener solution for Griffiths-Jim beamformers," *IEEE Trans. Signal Processing*, vol. 40, no. 2, pp. 474–478, 1992.
[22] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.
[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.
[25] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.
[26] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, NY, USA, 2nd edition, 2000.
[27] P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1987.
[28] T. F. Quatieri and R. Dunn, "Speech enhancement based on auditory spectral chance," in *Proc. 27th IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 257–260, Orlando, Fla, USA, May 2002.

Israel Cohen received the B.S. (summa cum laude), M.S., and Ph.D. degrees in electrical engineering in 1990, 1993, and 1998, respectively, all from the Technion – Israel Institute of Technology. From 1990 to 1998, he was a Research Scientist at RAFAEL research laboratories, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate at the Computer Science Department of Yale University, New Haven, Conn, USA. Since 2001, he has been a Senior Lecturer with the Electrical Engineering Department, Technion, Israel. His research interests are multichannel speech enhancement, image and multidimensional data processing, anomaly detection, and wavelet theory and applications.

Sharon Gannot received his B.S. degree (summa cum laude) from the Technion – Israel Institute of Technology, Israel in 1986 and the M.S. (cum laude) and Ph.D. degrees from Tel Aviv University, Tel Aviv, Israel in 1995 and 2000, respectively, all in electrical engineering. Between 1986 and 1993, he was the Head of a research and development section in R&D center of the Israel Defense Forces. In 2001, he held a postdoctoral position at the Department of Electrical Engineering (SISTA) at Katholieke Universiteit Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Signal and Image Processing Lab (SIPL), Faculty of Electrical Engineering, The Technion – Israel Institute of Technology, Israel. Currently, he is affiliated with the School of Engineering, Bar-Ilan University, Israel.

Baruch Berdugo received the B.S. (cum laude) and M.S. degrees in electrical engineering in 1978 and 1986, respectively, and the Ph.D. degree in biomedical engineering in 2001, all from the Technion – Israel Institute of Technology. From 1978 to 1982, he served in the Israeli Navy as an Engineer. From 1982 to 1997, he was a Research Scientist at RAFAEL research laboratories, Israel Ministry of Defense. From 1987 to 1997, he was Head of RAFAEL's R&D group of the acoustic product line. In 1998, he joined Lamar Signal Processing, Ltd. as a Vice President R&D, and since 2000, he has been the Chief Executive Officer. His research interests include multichannel speech enhancement and direction finding.